THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 2, February, 1972
Printed in U.S.A.

Man-Machine Interaction in Human-Face Identification

By A. J. GOLDSTEIN, L. D. HARMON, and A. B. LESK

(Manuscript received September 20, 1971)

How well can a computer identify a human face which is described by a person who is inspecting a photograph? We give an account of an interactive system that takes advantage both of the human's superiority in detecting and describing noteworthy features and of the machine's superiority in making decisions based on accurate knowledge of population statistics of stored face-features. Experiments using a population of 255 faces and 10 or fewer feature-descriptions showed that the population containing the described individual could be narrowed down to less than 4 percent in 99 percent of all trials.

1. INTRODUCTION

In a previous report[1] we described experiments in human-face recognition which were intended to establish a foundation for extended study. Those experiments provided a large body of reliable quantitative data based on 21 feature-descriptions of 255 human faces. These 21-dimensional vectors were shown to be sufficient for accurate individual identification, both by human and by computer search.*

The objective, then and now, is to explore new techniques for obtaining accurate recognition of vectors given imprecise component values. Our procedures involve searching through a population of vectors to retrieve one, a "target," whose components best match a searcher's imprecise specification.

There are two obvious kinds of such recognition and retrieval, just as in fingerprint-file search. One is that of finding the best match between an unidentified individual and a member of a file population. The other is that of assigning an individual to one of a number of

* Our population consisted of 255 white males aged 20–50 with no eyeglasses, facial hair, scars, or notable deformities. A panel of 10 observers independently evaluated 21 features (shown in Fig. 1) for each face. The average value of the observers' votes was used as the "official" description of each face-feature. Although individual feature-descriptions are restricted to *integral* values, averaging the panel's votes provides non-integral official descriptions. Reference 1 contains a detailed discussion of the features and population used.

predefined classes according to some systematic scheme. Ours is the first approach, matching, though the techniques developed could readily be used for the second, cataloging.

In our previous work, a subject was given a set of photographs of human faces and an official description of one of them. He was required to select that photograph which best matched the description. In the experiments reported here, the subject was shown a picture and was asked to describe it to a computer using features from a list given to him. The computer then searched a population of stored descriptions for best fit to the description furnished by the subject. In both studies we ran supplementary experiments employing computer simulation to establish theoretical limits of human performance under certain modeling assumptions.

In the earlier face-identification procedures, isolation was based on a binary-decision technique. At each step in the search, the population was progressively reduced by using a quantitative feature-description to determine which members of the remaining subset would be retained. On the average, eight feature-descriptions were required to isolate a face in a population of 255 males; about 50 percent correct identification was obtained. The binary process, however, obviously insured doom given just one error in the sequence.

A more lenient process is rank-ordering. If one ranks population members according to some goodness-of-fit criterion, any reasonably accurate description can be expected to place the target high on the rank-ordered list. Such a system can be quite useful in focusing attention on a small subset of the population that has high probability of containing the target. Population-reduction techniques like this are well-known to be useful in many tasks, from fingerprint-file search[2] and script recognition[3] to document retrieval.[4]

The present report deals with a real-time man-machine interactive system for human-face identification. The study has three main objectives:

(*i*) To develop a decision-making technique which replaces the earlier error-sensitive binary-decision selection process by a more forgiving rank-ordering process,

(*ii*) To design algorithms for optimizing the man-machine system so that we can take advantage of both the human's superiority in detecting noteworthy features and the machine's superiority in making decisions based on accurate knowledge of population statistics, and

(*iii*) To devise simple yet effective measures of performance.

II. SYSTEM DESIGN

The system design can be understood by considering our experimental procedure. A subject at a remote computer-terminal is given a photograph of one member of the population. He describes this target face to the computer using descriptive features chosen from a permitted set. The aim is to have the computer identify the target from the subject's description of it.

Subjects in our experiments used three-view photographs of target faces (two examples are shown in Fig. 6). The set of features from which descriptions were drawn is shown in Fig. 1.

In our experiments, features may be chosen by the subject or by the computer which uses an automatic feature-selection algorithm. There are three alternative modes of feature selection: the subject may choose all features, or he may choose some and then let the computer choose the rest, or the computer may choose all features.

After each feature description, the computer assigns a goodness-of-fit measure (a "weight") to each member of the population. This weight represents the similarity of the subject's description to the official description of each member of the population. At each feature-description step, the population is ranked by weight. After a predetermined number of steps, the process is terminated. We evaluate performance with respect to the target's rank and weight. An illustrative printout of one "portrait"* appears in Fig. 2.

Two aspects of system design are crucial: the weight-assignment algorithm and the feature-selection algorithm. They are described below. Following that, we discuss two critical experimental requirements, stopping criteria and measures of performance. The experiments reported in the succeeding section were designed to show how various modes of feature selection affected system performance.

2.1 *Weight Assignment*

The algorithm used to assign weights at each step must maintain a reasonable balance between penalizing descriptive errors so heavily that recovery from a mistake is impossible and penalizing these errors so lightly that no significant reduction of the population is achieved. The penalties assigned should distinguish between a minor descriptive error (e.g., medium-long vs long nose-length) from which recovery should be easy, and a major error (e.g., short vs long nose-length) from which recovery should be more difficult.

* A portrait is defined as a description consisting of a set of *integral* feature-values assigned by a subject; the subject is said to "portray" the target.

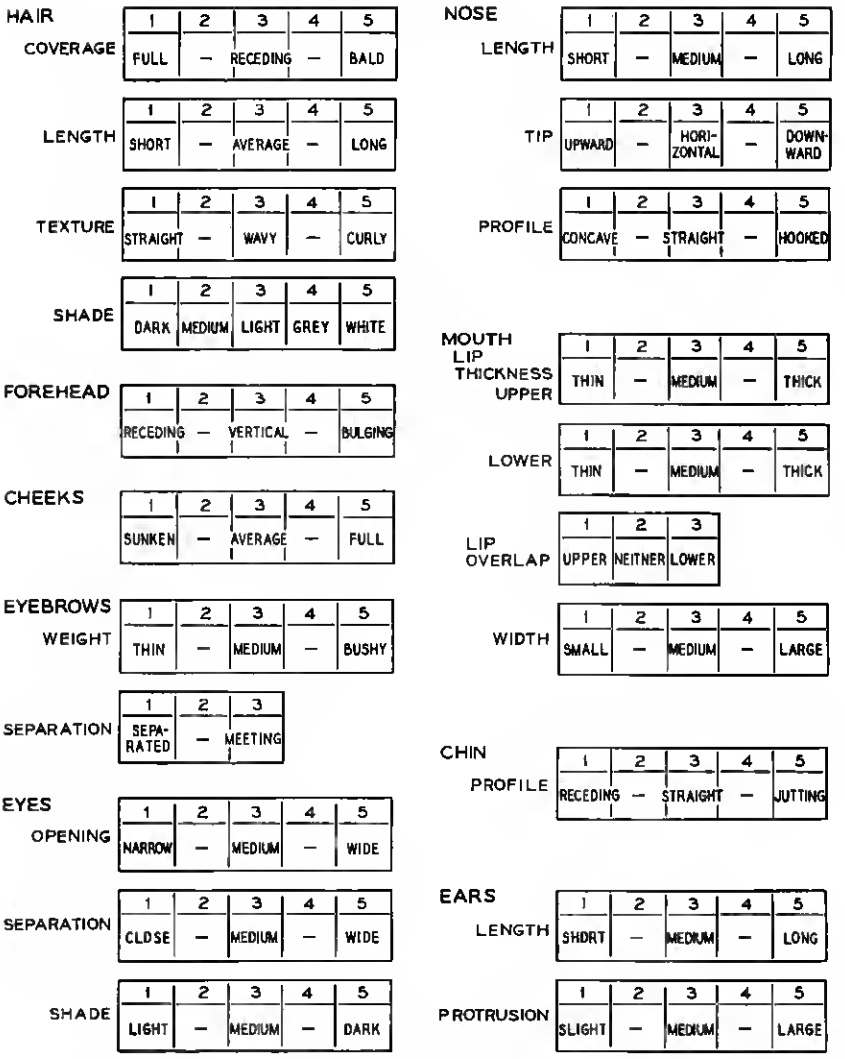

Fig. 1—Set of 21 face-features and their allowable values used for all experiments.

We chose

$$\sum_{i=1}^{s} |v_i - \hat{v}_i|^k = \sum_{i=1}^{s} \Delta_i^k$$

as the general form of an individual's weight at step s. For the feature described at step i, v_i is the individual's official value, $\hat{v}_i$ is the value

```
DESCRIBE NEXT PICTURE.

 FEATURE
       EYEBR0W WT.
 THIN        MEDIUM        BUSHY
   1     2     3     4     5
=1
   93  244  183  223  159
 1.00 1.00 1.00 1.00 0.82

 FEATURE
       EAR LENGTH
 SHORT       MEDIUM        LONG
   1     2     3     4     5
=1
   72  244  175   93   43
 1.00 1.00 0.82 0.67 0.66

 FEATURE
       LIP OVERLAP
 UPPER         NEITHER       LOWER
   1             2             3
=1
   72  226  114  122   76
 1.00 0.73 0.66 0.61 0.60

 FEATURE
       HAIR TEXTURE
 STRAIGHT      WAVY          CURLY
   1     2     3     4     5
=4
   76  122   32  244   52
 1.00 0.74 0.56 0.55 0.50

 FEATURE

        AUTOMATIC FEATURE SELECTION
 ******EYE SHADE
 LIGHT       MEDIUM        DARK
   1     2     3     4     5
=3
   76   52   72  221  191
 1.00 0.56 0.45 0.38 0.36

 ******EYEBR0W SEP.
 SEPARATE    MEDIUM        MEETING
   1             2             3
=2
   76  147   52   84   72
 1.00 0.50 0.42 0.37 0.34

 ******EYE OPENING
 NARROW      MEDIUM        WIDE
   1     2     3     4     5
=2
   76   72  226   26  191
 1.00 0.51 0.40 0.38 0.36

 ******UPPER LIP
 THIN        MEDIUM        THICK
   1     2     3     4     5
=3
   76  191   72  221   52
 1.00 0.33 0.28 0.23 0.21

 ******HAIR SHADE
 DARK  MED.  LT.    GRAY  WHT.
   1     2     3     4     5
=2
   76  221   72  226  191
 1.00 0.34 0.34 0.33 0.25

 ******LOWER LIP
 THIN        MEDIUM        THICK
   1     2     3     4     5
=1
   76   72  221   84  191
 1.00 0.19 0.13 0.12 0.11

 PLEASE TYPE TARGET NUMBER.
=76

 ORDER FEATURE          DESCRIPTION      RANK
                        YOU    AVG.   NO.    %
   1   EYEBR0W WT.       1     2.2    27   10.2
   2   EAR LENGTH        1     2.3     8    2.7
   3   LIP OVERLAP       1     1.2     5    1.6
   4   HAIR TEXTURE      4     3.0     1    0.

   5   EYE SHADE         3     2.7     1    0.
   6   EYEBR0W SEP.      2     1.3     1    0.
   7   EYE OPENING       2     2.6     1    0.
   8   UPPER LIP         3     2.9     1    0.
   9   HAIR SHADE        2     1.5     1    0.
  10   LOWER LIP         1     2.3     1    0.
```

Fig. 2—Printout of one interactive dialog. Computer requested feature; subject picked *Eyebrow Weight*. Computer printed allowable feature-values; subject voted *thin*. In next two lines computer displayed calculated weights of the top five individuals. First four faces, 93 · · · 223, tied with relative weights 1.00. Face no. 159 in fifth place was weighted 0.82 relatively. By step three the target (no. 76) was in fifth place, advancing to first rank by step four despite deliberately introduced errors on first two steps. Subject changed to AFS at step five, whereupon computer specified *Eye Shade*. Nearest neighbors were gradually separated; by step 10 the closest had relative weight of only 0.19. Portrait automatically terminated at step 10. Summary compares subject's assignments with official values ("AVG."); also displayed is target's rank at each step and percentage of population with higher rank.

assigned by the subject, and Δ_i is the magnitude of the difference between them.

A number of variants of this formulation were tested. In particular we found that $k = 1$ yielded results as good as or better than any other value of k. We also considered the effect of quantization error

arising from comparing the integral feature-values used by subjects to the non-integral official feature-values. For each feature description this error is at most 0.5. Alternative formulations of weight functions intended to minimize the effect of this error degraded performance. Our earliest formulations of weight used an exponential form. While presently unessential, the exponential has survived in our computational algorithms. Consequently, with $k = 1$ and with no compensation for quantizing effects, the weight assignment is

$$W = \exp\left(-\sum_{i=1}^{s} \Delta_i\right).$$

2.2 *Automatic Feature-Selection*

As noted above, features may be selected either by the subject or by the computer. The two methods have complementary advantages: The subject possesses exhaustive knowledge of the face he is portraying, but he knows very little about the characteristics of the population stored in the machine; conversely, the machine does not know who the target is, but it does possess the official descriptions of all population members and their goodnesses-of-fit to the target description.

We wish to find if the advantages of human and of computer feature-selection can be usefully combined, where the human can take advantage of *extreme* features, while the computer can utilize *discriminating* features.

An *extreme* feature of a target is a feature whose official value is near an extreme of that feature's range; e.g., long hair, short nose, small mouth. This classification does not depend on the target's other feature values or those of the population. It depends only on the feature's value and range.

Conversely, a *discriminating* feature is a purely relative concept, based on the population and the target description up to any given step. At each step, we refer to a feature as discriminating if its description will distinguish among those individuals whose official descriptions match the partial portrait well (i.e., the individuals who have large weights). Whether a feature is discriminating depends on the statistics of feature-value distribution over the population.

We wish to develop an automatic-feature-selection procedure that chooses the most discriminating feature available as the next one to be described in a portrait. How can we decide when a feature is discriminating?

Consider the two hypothetical distributions of official feature-values

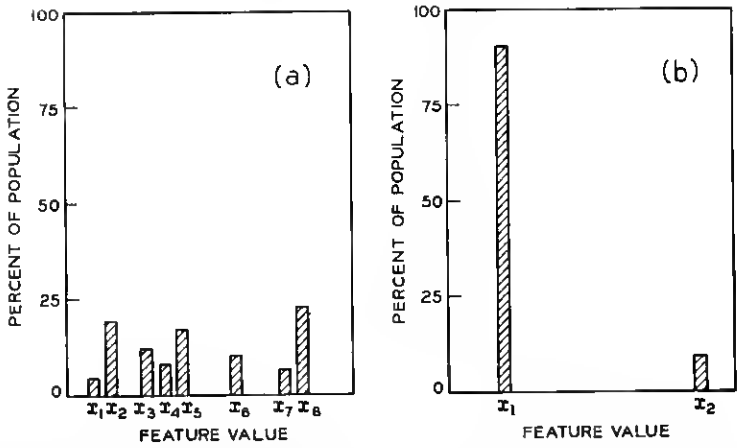

Fig. 3—Two types of distributions of official feature-values: (a) Relatively uniform distribution represents *discriminating* feature. (b) Relatively non-uniform distribution indicates a *nondiscriminating* feature.

shown in Fig. 3. If feature b were used, and if the target's value happened to be x_2, then the target would be well separated from the rest of the population. It is much more likely, however, that the target's value would be x_1, in which case the separation of population members would be poor. If feature a were used, one would always obtain some intermediate amount of population separation. In the extreme, if all members of the population had the same value of a particular feature, say very long ears, then the use of that feature would not lead to population separation. Conversely if the values were uniformly distributed over the population, maximum discrimination and most effective separation would be obtained.

In considering feature-value distributions, it is undesirable to utilize the official description of *every* member of the population for all unused features. Not only would this increase cost, but it would degrade performance. This can be seen from the following argument: The aim of automatic feature-selection (AFS) is to find a feature which will decrease the number of individuals who are described well by the portrait thus far. The distribution of feature values among those individuals may be completely different from the distribution in the whole population. If AFS considered all individuals, the distinguishing characteristics of the high-ranking individuals would be obscured by those of the overwhelming number of low-ranking individuals. To avoid waste of one's knowledge of the partial portrait, AFS considers the distribution of feature values only in the subset of the population

which the portrait describes well, although the feature chosen will be used to rerank the *entire* population. This subset should include the individuals who could easily attain first place in the rank-ordered population. In practice, we found that to consider all individuals with weight ≥ 0.7 times the weight of the first-ranked one, but at least 10 individuals, was effective.

As a result of the above arguments, we implemented an AFS procedure which chooses as the next feature that one for which the distribution of the feature-values of the high-ranking individuals is most nearly uniform. This will be the most-discriminating feature in the sense of efficient identification. Analytical details of the procedure are given in Appendix A.

2.3 *Stopping Criteria*

The portrait composition must continue for enough steps to insure accuracy. On the other hand, too many steps lead to subject fatigue and boredom. The rule which governs when portrait composition stops should satisfy both these requirements.

A stopping rule may be *dynamic* and depend on the ranks and/or the weights at each step, or the rule may be *static*, e.g., stop after a predetermined step. Our earlier experiments, employing a human binary-search process,[1] showed that, on the average, fewer than eight features were used when a target was successfully identified. One might conjecture that with 5-valued features some 2.3 bits of information could be available at each step, and so the present experiments should require fewer than 8 steps for isolation, and not less than $\log_2 255/\log_2 5 = 3.5$.

This argument, and information from trial runs indicating that fatigue and boredom commenced after the subject judged about ten features, were used to arrive at a static stopping-rule of ten steps. Experimental results have shown this to provide adequate accuracy. The data we obtained permitted us to formulate an efficient dynamic stopping-rule for future use; it is described in Section IV.

2.4 *Measures of Performance*

A binary search-procedure may be evaluated by whether and at what stage the target is ultimately isolated, or at what stage the target is rejected and the size of the smallest subset that contained the target. Meaningful measures of performance for a rank-ordering procedure are less obvious.

One useful measure, population reduction, can be transferred directly from binary search to rank ordering. We can consider the size of the

subset of the population with rank greater than tbat of tbe target, and how rapidly the population is reduced to that size. The concept of absolute isolation is thus replaced by one of relative identification.

We measure the population reduction at each step by the rank of the target. Since his rank usually changes from step to step, we use as an overall measure of performance the mean rank of the target from the sixth through tenth steps. The first five steps are not included because the target's rank then is usually large and changing rapidly.

Population reduction shows whether the target is separated from the rest of the population. It does not reveal, however, the *extent* of that separation. To do this, a "confidence" measure was introduced. It is based on the weights of the individuals in the ranked list, as follows: If the target is ranked first, his confidence is the ratio of his weight to that of the second-ranked individual; otherwise, the target's confidence is equal to the ratio of his weight to that of the first-ranked individual. A confidence value less than 1.0 denotes failure to place the target in first rank; confidence values greater than 1.0 correspond to varying degrees of success. Obviously, the magnitude of the confidence measure depends on the weighting function being used.

Confidence and rank are useful in evaluating a single portrait; their averages can be used to compare several sets of portraits. A third measure we find useful is the *rank cross-section*; this is meaningful only for comparing sets of portraits. For a set of portraits, the rank cross-section is the frequency with which targets reach or exceed a given threshold rank (e.g., first rank, or top 2 percent of population, etc.) at each step of a portrait. This indicates the average speed and extent of a target's rise in rank.

However, a target does not necessarily always rise in rank. A faulty feature-judgment may worsen his position. The weighting scheme is forgiving in that it permits recovery from a subject's error in feature judgment. Another way of viewing this is that once the target is entrenched in first place, i.e., has a large confidence, it takes a large error in judgment to displace him.

We can express this quantitatively as follows: Suppose the target is in first rank; let him have confidence c, and let the next feature judgment for him have an error Δ. Suppose that the error for the second-ranked individual is 0. Then with the weighting scheme that was adopted, we find that if $\Delta > \ln c$, the ranks will be reversed. Thus,

when confidence $c \leqq$	1.6	2.7	4.5	7.4	12.2	20.0
reversal occurs if $\Delta >$	0.5	1.0	1.5	2.0	2.5	3.0.

Data on subject error (see Section 3.1.1.1) show that 95 percent of the time $\Delta \leqq 1.0$. Thus a first-ranked target with confidence 2.7 or greater is rarely dislodged.

III. EXPERIMENTS

3.1 *Human Experiments*

An interactive experiment was run to evaluate the effectiveness of our overall system and to test the relative utility of three different modes of operation.* In one mode the subject selects *every* feature he describes to the computer, using first those he considers most extreme for the target. We shall refer to this mode as "NO AFS" (i.e., no automatic feature-selection). In another mode, termed "ALL AFS," the subject simply assigns feature values for each feature specified by the computer which is operating in the automatic-feature-selection mode described earlier. A third mode, termed "MIXED," requires a subject to select features until he decides there are no more he considers outstanding, then to invoke AFS.

We expected subject selection of extreme features to enhance separation, at least for the first few features, for many members of the population. When there are no extreme features to use, then computer selection of discriminating features should facilitate target separation. We expected that the mixed mode of operation, taking advantage of the best capability of both human and computer, would yield best results as measured by confidence and rank.

Fifteen subjects were used (13M, 2F). Twenty-one features were made available, as illustrated in Fig. 1. Each subject participated in three separate sessions, one in the NO-AFS mode, one in MIXED, and one in ALL AFS. Each of the 15 subjects, portraying 15 targets, provided us with 225 portraits. Five targets were portrayed in each session. Fifteen different targets were used; each subject thus portrayed all targets. The targets were individually selected at random from our population of 255; as an ensemble they were shown to preserve the feature distributions of the entire population. To minimize possible effects of learning, we randomized the order in which subjects used the three modes of feature selection and the order in which they portrayed the targets.

At the beginning of the experiment each subject was given 20–30 minutes of verbal instruction to familiarize him with the feature set. This used a collection of sample faces that were not employed in the

* The program which was used is described in Ref. 5.

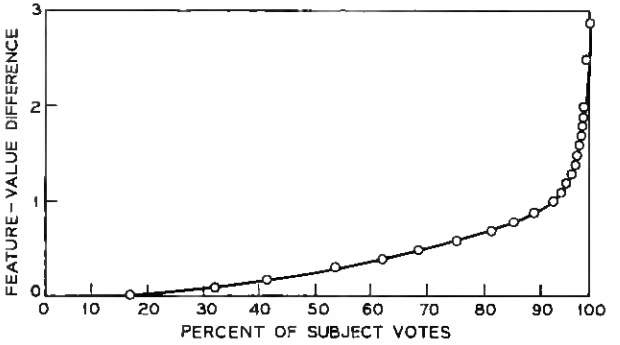

Fig. 4—Cumulative distribution of differences between subject votes and official values. The difference was never greater than 1.0 for 95 percent of the votes.

experiment. The subject then observed the experimenter portraying one target.

At the beginning of each session, the subject portrayed one practice target using the same mode of description (NO AFS, MIXED, or ALL AFS) to be employed in the experimental session. In all cases the subjects viewed the target's photograph while describing his features.

3.1.1 *Results*

3.1.1.1 *Feature-Judgment Reliability.* Our 15 subjects, making 2250 total judgments (15 subjects × 15 targets × 10 features), were in excellent agreement with the official feature-values. This can be seen in Fig. 4 which displays the cumulative distribution of magnitudes of the differences (Δ) between the subject judgments and the official values. In 95 percent of the 2250 judgments, the Δ was $\leqq 1.0$ (the maximum Δ is 4.0 for a 5-valued feature).* No judgments were as much as 3.0 off, only two were > 2.0 off, and only 24 of the 2250 judgments were different from the official values by more than 1.5.

Standard deviations were computed for the distributions of subject judgments, feature by feature. In both the ALL-AFS and the NO-AFS experiments, the standard deviation ranged from 0.42 to 1.1. The standard-deviation values for each feature are similar for ALL AFS and NO AFS, indicating no significant difference in subject accuracy as a function of whether feature selection is active or passive.

3.1.1.2 *Identification Accuracy.* The confidence and rank data,

* With the exception of two three-valued features. The data of Fig. 4, which include all 21 features, are not significantly changed by deleting the contributions of the two three-valued features.

averaged over all subjects and all targets, are shown in Fig. 5. For the combined 225 portraits, the mean confidence at step 10 was 5.65, and the mean rank over the sixth through tenth steps was 4.12. For 75 MIXED portraits, the mean confidence and rank were 6.79 and 2.75 respectively, while for 75 ALL-AFS portraits the corresponding figures were 4.41 and 6.71. The results of the 75 NO-AFS experiments were intermediate; mean confidence was 5.74, and mean rank was 2.91.

Subject performance varied considerably. Both the average confidence and the average rank had a range of 6:1 (from best to worst subjects). One subject's performance was consistently poor. When his scores are deleted, the average rank improves from 4.12 to 3.70, and the average confidence improves from 5.65 to 5.80.

To test for improved performance with practice during the course of the experiment, the data for each subject were examined according to their temporal sequence. No trends were observed.

The 15 targets received a rather wide range of performance indices. Number 99 had an average confidence measure of 20.3 (compared to the 15-target mean of 5.65), and his average rank was 1.39 (compared to the 15-target mean of 4.12). At the other extreme, no. 19 had a confidence measure of 0.88 and a rank of 9.16. These two individuals are depicted in Fig. 6.

	NO AFS	MIXED	ALL AFS	COMBINED
CONFIDENCE	5.74	6.79	4.41	5.65
RANK	2.91	2.75	6.71	4.12

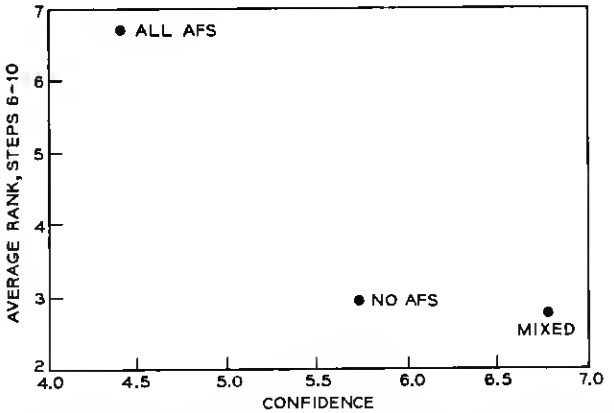

Fig. 5—Two measures of performance summarized for all subjects and targets. MIXED mode is clearly superior, while ALL AFS is markedly poorest. Combined results for all experimental data show that the average target, with a rank of 4.12, was in the upper 1.6 percent of the population over the sixth through tenth steps.

NO. 19

NO. 99

Fig. 6—Targets which produced two extremes of performance. No. 19 was difficult to retrieve, obtaining confidence 0.88 and rank 9.16, while no. 99 was outstandingly easy, obtaining confidence 20.3 and rank 1.39.

The reasons for the different success with the two targets are clear. In general, no. 19 is much closer to the population mean than is no. 99 who has a larger number of more extreme features than has no. 19. All ten subjects who portrayed no. 99 in either the MIXED or the NO-AFS mode started their portrait with hair texture; no. 99 has the curliest hair in the population. All ten also described his light hair-shade and thin upper lip, and all five NO-AFS portraits included his small-to-medium mouth width. By contrast, only one of no. 19's features received unanimous mention: his medium-to-wide eye opening.

3.1.1.3 *Performance Differences Among NO AFS, MIXED, and ALL AFS.* The differences in performance among the three modes of opera-

tion are clear and consistent. This can be seen first by noting the average rank of the target at each feature step. Figure 7 illustrates this by a plot of the percent of the population with better rank than the target at each step. Overall, the population reduction in early steps is quite rapid.

It is clear that at any step the ALL-AFS mode places the target about twice as far down the ordered list as does either of the other two modes. This suggests that knowledge of the population statistics is not as effective as knowledge of a target's outstanding features. Both the MIXED and the NO-AFS modes are roughly equal and are superior to ALL AFS. From step seven on, with the MIXED and NO-AFS modes, the population having better rank than the target was reduced to 0.68 percent. We have seen (Fig. 5) that the confidence in the MIXED experiments is 18 percent higher than that in the NO-AFS experiments and 54 percent higher than that in the ALL-AFS experiments. Similarly, the rank results are superior for MIXED, being 11 percent ahead of NO AFS and 59 percent ahead of ALL AFS. Even for ALL AFS, however, the average rank was better than seventh place; i.e., 2.2 percent of the population had better rank than the target.

The plots of rank cross-section (see Section 2.4), displayed in Fig. 8, also make evident the relative inferiority of ALL AFS. The asymptotic levels of NO AFS and MIXED are virtually identical. For both MIXED and NO AFS, half the targets reach first place by step five, and by

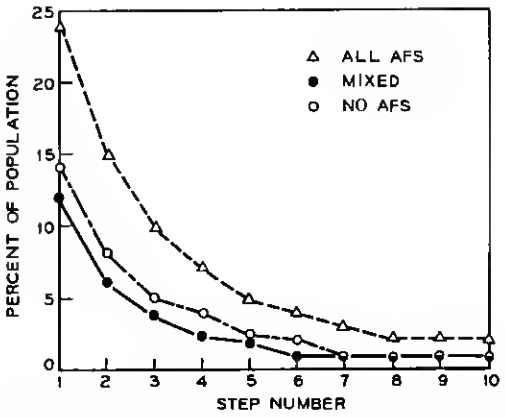

Fig. 7—Comparison of how three modes of system operation affect the percent of the population having better rank than the target. MIXED mode is clearly superior in early steps; with eight feature-steps ALL AFS reduces the population to 2 percent, and both other modes reduce it to 0.68 percent.

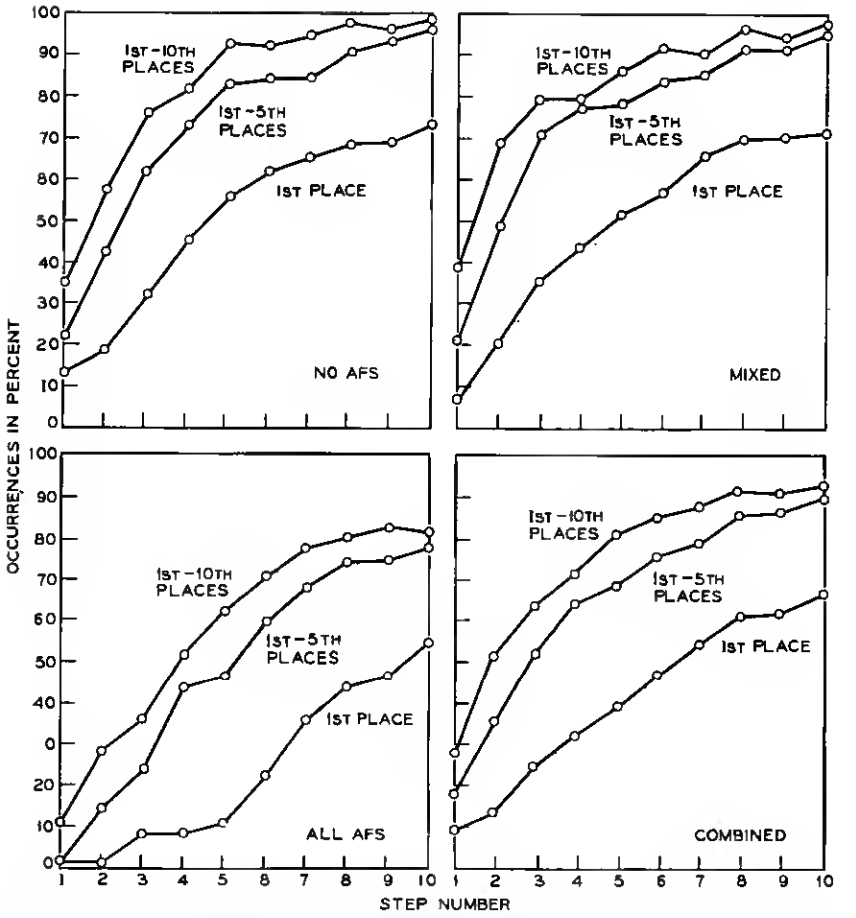

Fig. 8—Rank cross-section at each step. ALL-AFS mode is obviously inferior. Performances from NO AFS and MIXED are essentially alike. By step five, roughly half the targets reached the top in MIXED and NO AFS; by step 10, better than 70 percent reached first place.

step ten in both modes 99 percent of the targets are in no worse than tenth rank. And 96 percent are in no worse than fifth rank.

Although ALL AFS does not produce results comparable to those of the other modes, more than half the targets reach first place by the tenth step, and more than three-quarters of them reach fifth place or better.

The confidence measure (see Fig. 5) also indicates the relative inferiority of ALL AFS. Unlike the other measures discussed here, confidence

shows MIXED to be superior to NO AFS in separating the target from the rest of the population.

3.2 *Computer Experiments*

How does human performance compare with that of an "ideal" subject? The major variables in subject performance are the set of features selected, the accuracy with which they are judged, and the order in which they are described. Since the subject is constrained to use integral feature-values, the best judgment he can make on any feature is the nearest integer to the target's official description; we shall refer to this value as a "rounded" judgment. For each target there is a sequence of features which gives the largest confidence at step 10, and there is one which gives the best average rank. Either of these could be regarded as the optimal sequence chosen by an ideal subject. However, there is no easy way to find such optimal sequences; therefore the ideal subject was defined as follows:

For each target the sequence of features to be used by the ideal subject in a computer simulation was selected on the basis of feature "extremeness." The extremeness of an individual's feature is the magnitude of the difference between his official value and the feature's population mean. Our ideal subject, modeled on how our human subjects were instructed, was defined to be one who selected features in descending order of extremeness *and* used, for each feature's value, the rounded value of the official description.

This ideal subject was used to portray the 15 targets employed in the human experiments. The distribution of the step at which the target first achieved rank one and remained there through step 10 is

Step	1	2	3	4	5	6	7	8	9	10
Frequency	2	4	4	3	1	0	0	0	0	1.

For all targets, the average number of steps is 3.27, and the average rank (over steps 6 through 10) is 1.01 (i.e., virtually perfect). The confidence at step 10 ranged from 1.00 to 95.6 with an average of 21.5 and a median of 16.1.

These results are markedly superior to the results of the human experiments summarized in Fig. 5. Are the differences due to subject judgment-errors or to less-than-ideal feature selection owing to the fact that the subject does not know the population statistics?

To explore this question, three additional computer studies were performed with the same 15 targets used in the human experiments. The results of all four computer experiments are summarized in the tabulation below and are contrasted with the NO-AFS human experi-

ments. Experiment no. 1 is that described above, using the ideal subject. In the second experiment, the NO-AFS human experimental data were modified by replacing the subject judgments with rounded official-values. Third, the extreme features chosen by the ideal subject were used with human judgments. In the last computer experiment (no. 4), four random sequences of features were used with rounded feature-judgments. Finally, the results of the NO-AFS human experiments are shown.

Besides displaying confidence and mean rank (averaged over steps 6 through 10), the table shows the number of targets on which confidence was greater than, approximately equal to, and less than the confidence obtained by the ideal subject.

Exp.	Feature Selection	Judgment	Conf.	>	≅	<	Mean Rank
1	Extreme	Rounded	21.5	0	15	0	1.01
2	NO AFS	Rounded	10.6	4	6	5	1.23
3	Extreme	Human	8.25	1	2	12	1.68
4	Random	Rounded	4.08	0	2	13	1.76
5	NO AFS	Human	5.74	2	0	13	2.91

The confidence and mean rank show the performance of the ideal subject (exp. no. 1) to be better than that obtained in the experiment using NO AFS and rounded official-values. Notice, however, if one examines confidence for the ideal case and NO AFS rounded, *target by target*, then it is seen that NO AFS is better about as many times as it is worse. Since the only variable was feature selection, this indicates that the humans were almost as good as the ideal subject in their choice of features. The use of extreme features with human judgments (exp. no. 3) gives worse performance in rank and confidence than does NO AFS with rounded judgments. This shows that the advantage of extreme-feature selection was not sufficient to overcome human errors in judgment.

It might be argued that *any* feature sequence would produce good results. But the random experiment shows that perfect feature-judgments alone are not sufficient; feature selection is important.

In summary, humans are nearly ideal in feature selection while considerably less than ideal in feature-value assignment.

IV. EXTENSION TO LARGE POPULATIONS AND TO OTHER PROBLEMS

The procedures we have described for identification and retrieval are applicable to problems other than the face-recognition tasks we have

so far explored. Such searches as medical diagnosis and telephone-directory lookup also deal often with noisy data where probabilistic identification is made. With what generality can the procedures we have evolved be applied to tasks where descriptive components are imprecise and populations are large?

First, however, there are questions of economic feasibility. The storage and computing requirements in the present experiments are modest. For a population of 255, we require 1500 words* of disk and 14,400 words of core storage. Memory requirements grow at a rate of 7 words/face. The interactive computation process (slowed enormously by the human at a remote terminal) takes about 5–10 minutes real time (~5 seconds central-processor time) and costs $2.50 on the average. A key question for extended applications is: How do these numbers increase with population size?

In the earlier model of the binary-search identification process,[1] we showed a logarithmic growth of the number of steps (features) required to isolate a target. For a particular condition we found useful, the model predicts that an average of only 13.5 feature-descriptions will be required for a population of 4 million. If the actual growth of the number of steps required to isolate in the present rather different rank-ordering process is close to our model's prediction in the binary-search process, then a nonlinearity very important to economic treatment of large populations will be at hand. That this may indeed be so can be seen in Appendix B.

To investigate the effect of population size on the number of steps required for isolation, comparable runs were made with population sizes of 128, 255, and 510 individuals.† The first feature in all portraits was chosen at random, and all subsequent features were chosen by AFS. (Since the number of individuals used in the AFS computation is a function of each partial portrait, the cost varies from target to target.) The dynamic stopping-rule described at the end of this section was used. Feature judgments were drawn from the panel of observers whose averaged judgments comprise the official values. Randomly chosen observers supplied portraits. The data for each population size were averaged over five portraits of each of 15 randomly chosen individuals (75 portraits total). The results of this experiment are summarized below.

* The computer is a time-shared Honeywell-635 having 36-bit words.

† The 128-individual population is a randomly-chosen subset of the 255-face one. The 510-individual population is composed of the original 255 individuals plus 255 "new" pseudo-faces created by randomly shuffling the feature values of the old population.

	Population Size		
	128	255	510
Mean stopping step, std. dev.	9.5, 3.9	10.6, 4.2	11.7, 4.3
Relative total cost	1.00	1.98	4.56
Relative cost/step	1.00	1.77	3.76

While the mean stopping step appears to increase logarithmically with population size, P, the cost per step increases roughly in proportion to population size. That is,

$$\text{Total Cost/Step} \sim P$$

and the logarithmic growth of the mean stopping step with population size gives

$$\text{Total Cost} \sim P \ln P.$$

The mean stopping step increased very slowly with population size, from 9.5 to 11.7 for populations of 128 and 510. The final rank of the target rose on the average from 1.4 to only 2.5, less than a twofold increase for a fourfold increase in population size. Experience with MIXED and ALL AFS indicates that the corresponding figures for MIXED would be markedly better than those above, which were obtained with ALL AFS.

The cost of the AFS algorithm is linear with respect to the number of faces used to determine the next feature. Figure 9 shows that this number converges rapidly to a minimum. It is seen that, at most, less than 35 percent of the population is used in the AFS computation at step two and less than 15 percent at step three. From step four on (with but a slight exception at step five), only 3.9 percent is used; this is the minimum possible given our (arbitrary) convention of considering all individuals with relative weight ≥ 0.7, but at least 10 faces (10/255 = 3.9 percent).

Several kinds of algorithmic corner-cutting look attractive and are under consideration. The results displayed in Fig. 10 show that for a given performance level only some minimum proportion of the population need be considered at each step. For example, if flawless performance were required while operating in the MIXED mode, no more than half the population would need to be considered in steps three and four, and from step five on, at least 75 percent of the population could be ignored. In 95 percent of all trials, the target was in the top 10 percent of the population from the sixth step on. The computational

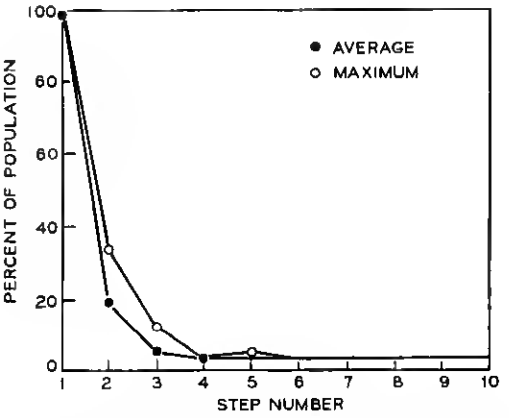

Fig. 9—Extent of AFS computation. For empirically determined rule of considering only those population members with weight equal to or greater than 0.7, or a minimum of 10, extent of computation drops rapidly. With but very slight exception, at and after step four no more than 10 individuals have weights above 0.7, indicating efficient separation of top members.

savings with such a limited-depth search would thus be considerable.

Another possible economy might be some form of individual or feature clustering. One could divide the population into small groups of "look-alikes" and create a "super-description" for each cluster whose official description was the mean of the individual descriptions. One could then order these clusters according to their resemblance to the target description and then search the clusters' members in that order to find a good individual match to the description. This scheme assumes that such a clustering can be achieved and that the cluster descriptions would be non-trivially different.

In a sense the 255 individuals we have dealt with comprise a cluster of the general population. Our 255-member subpopulation was deliberately chosen to be homogeneous (see footnote on page 399) to make isolation more difficult. Consequently, several highly reliable features (e.g., gender, race, age) could be added to our feature set for use with a more universal population. We might guess that the general population represented by the nonrepresentative subpopulation used in these studies is on the order of several thousand individuals.

An Empirical Dynamic Stopping Rule

An empirical dynamic stopping rule was developed using the data gathered from the 75 NO-AFS portraits. It is based on the concepts of confidence and rank and on tradeoff between the frequency and accuracy with which the rule stops portrait composition.

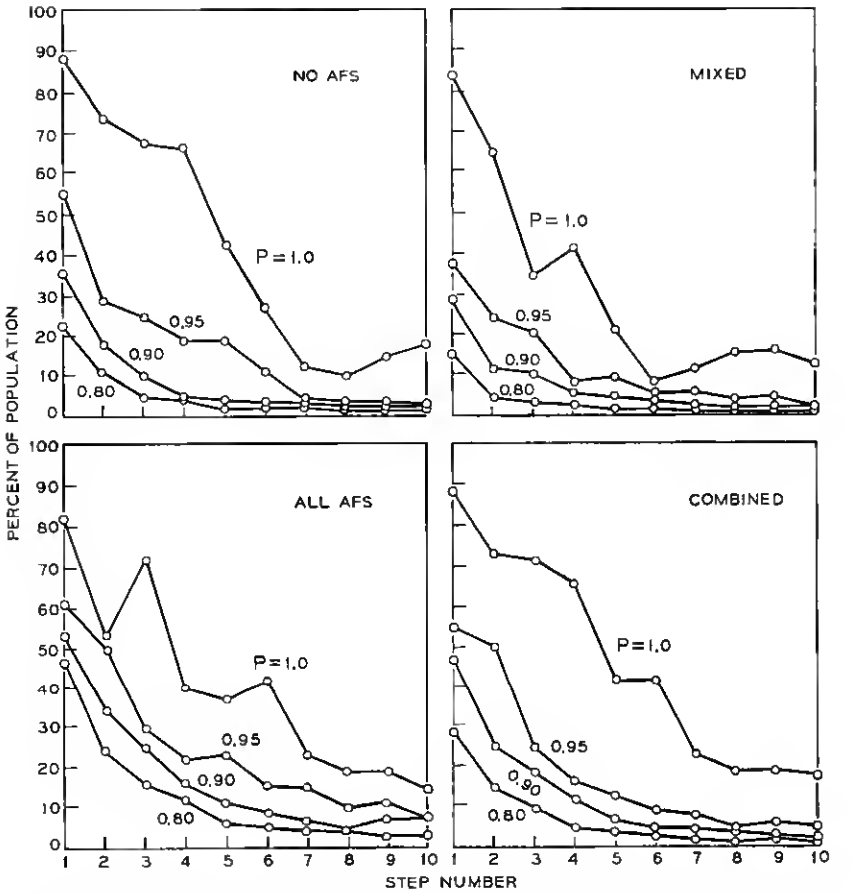

Fig. 10—Minimum envelope needed to capture the target with several probabilities at each step. $P = 1.0$ corresponds to the *worst* rank observed experimentally. After step five, target was in top 10 percent of the population for all cases except ALL AFS.

We consider first the *confidence*, which measures the degree of separation among population members. To formulate a stopping rule, we will use a variant "pseudo-confidence," the ratio of the weights of the first- and second-ranked individuals. (Note that this ratio is always $\geqq 1.0$). The experimental data show that when this ratio exceeded 3.5 at any step in the portrait, the target was then ranked first in 32 of the 34 cases, and the first-ranked individual was subsequently unseated in only two of 34 cases. We adopt this threshold as one component in our dynamic stopping rule: Whenever the pseudo-confidence exceeds 3.5, stop portrait composition.

Unfortunately, such a high pseudo-confidence occurs in fewer than half of the portraits. Another possible stopping criterion is an extended tenure-of-first-place by the same individual. Consequently, we adopt as the second component in our stopping rule: If the same individual has been in first place for the last six steps, stop regardless of the value of the pseudo-confidence. In only one of 37 cases did an individual change rank after holding first place for six or more consecutive steps.

We now have two criteria which would have terminated 80 percent of our experimental NO-AFS portraits. It was decided to use them as points on a linear stopping rule combining p, the pseudo-confidence, and s, the number of consecutive steps in which the same individual has been first-ranked: If $s + 2p > 8$, stop. This is the dynamic stopping rule used above to compare costs for various population sizes.

This empirical stopping rule was applied to the data from the rest of the experiment, and it provided another means of comparison (the mean stopping step) among the three types of portraits. The table below shows the results of applying the dynamic stopping rule to the NO-AFS, MIXED, and ALL-AFS runs.

	NO AFS	MIXED	ALL AFS
Decisions (Number of portraits terminated by stopping rule)	55	56	43
Correct decisions	49 (89%)	48 (86%)	30 (70%)
Mean stopping step, std. dev.			
Decisions only	6.8, 2.1	6.6, 2.2	7.7, 1.8
All portraits	7.7, 2.3	7.4, 2.4	8.7, 1.8
Mean rank of target			
Decisions only	1.4	1.6	3.6
All portraits	2.3	2.4	5.1

The number of decisions is the number of portraits (out of 75 in each case) which met the requirements of our stopping rule at or before the tenth step. A correct decision is one in which the target was in first place at the stopping step. The mean stopping step and its standard deviation are given for both the portraits which the stopping rule terminated ("Decisions only") and for all 75 portraits, considering the stopping step to be 10 for portraits in which no decision was made. The mean rank of the target at the stopping step is also given for both cases.

The data show the performance of MIXED and NO AFS to be almost identical. Both are superior in all respects to ALL AFS. The

mean stopping step and mean rank of the target are in the ranges one would expect from Fig. 7, which shows the progression of the average rank of the target. The stopping rule usually was satisfied soon after the position of the target had stabilized.

If this dynamic stopping rule had been used in our experiments, the average stopping step for a portrait would have been 7.9 instead of 10, a 21-percent saving with virtually no loss of accuracy in identification.

V. SUMMARY

An interactive system for the description and retrieval of multi-dimensional objects has been developed. This paper describes the system and its performance in face-identification experiments.

The system permits flexible description of target items using features chosen by either the user of the program or an automatic-feature-selection algorithm. At each step, AFS selects the feature which is most likely to be discriminating. It makes this choice on the basis of the partial portrait and the population statistics. Population members are ranked at each step on the basis of weights which reflect the match between the portrait description and each individual's official value. Performance is measured by two indices, confidence and rank.

The system was evaluated using 21 features, a population of 255 faces, and three modes of operation (NO AFS, MIXED, and ALL AFS). There were four principal results:

(*i*) The population was quickly and effectively reduced by all modes of operation. Over all trials, the population was reduced to less than 4 percent more than 93 percent of the time, and the target was successfully "isolated" (i.e., was in first place by portrait's end) 67 percent of the time (see Fig. 8). In 95 percent of all trials, the target remained in the top 10 percent of the population from the sixth step on.

(*ii*) The MIXED mode was the most effective in separating the target from the rest of the population as measured by confidence (see Fig. 5).

(*iii*) MIXED and NO AFS were equally effective with respect to population reduction, as measured by rank. The performance of these two modes was considerably superior to that of ALL AFS (see Figs. 5, 7). In the MIXED experiments, the population was reduced to less than 4 percent over 99 percent of the time, and the target was isolated 70 percent of the time (see Fig. 8).

(*iv*) The extent of the AFS computation drops rapidly with step number, reaching its minimum by step four (see Fig. 9).

These results can be summarized as follows: even in the worst case there is fair performance in singling out a target and good performance in narrowing down the population; and in the best case the population reduction is excellent.

This rapid population-reduction and the slow growth of the mean stopping step with population size (using the dynamic stopping rule) make the extension of these experiments to larger populations feasible. To process very large populations, say on the order of a million, new approaches would undoubtedly be needed. With the cost-cutting modifications we have described (dynamic stopping rule, limited-depth search), the present system could economically accommodate a population on the order of 5000.

VI. ACKNOWLEDGMENTS

We appreciate the skill and the stamina both of our subjects and of our early-draft critics W. S. Brown, Murray Eden, E. N. Gilbert, Newman Guttman, M. E. Harmon, S. C. Johnson, M. E. Lesk, R. C. Lummis, David Slepian, and Eric Wolman.

APPENDIX A

Automatic Feature-Selection

As discussed in the text (Section 2.2), the automatic-feature-selection algorithm selects, at each step, the most discriminating feature for the subject to describe next. The purpose of this Appendix is to formalize what is meant by a discriminating feature.

The AFS algorithm uses a subset of the population whose members are well-described by the subject's description of the target. In order to give greater importance to those members of the subset with high weight, each member's official feature-values were considered in proportion to his weight. The most discriminating feature, for that subset, thus is the one for which the distribution of the weighted feature-values is most uniform. Since the distribution of feature values may span different parts of the permissible feature ranges, distributions are shifted to facilitate equitable comparisons among features.

We shall define, for any shift, the deviation of the distribution of weighted feature-values from a uniform distribution. Formulae for the best shift and corresponding deviation are then derived.

Consider the subset of the population whose members are well-described by the subject's description of the target. Let the members of this subset have weights $W_1, \cdots, W_n$. The sum of the weights is W_T.

Let us concentrate on one feature. For convenience, scale its range to be from 0 to 1. Let the (scaled) official values corresponding to the above weights be $v_1, \cdots, v_n$.

Let

$$p_i = W_i/W_T .$$

We may interpret p_i as the probability that individual i is the target. When the sum of these probabilities in any interval is equal to the length of that interval, then the distribution of weighted feature-values is uniform. That is, if the interval is (x_1, x_2), then

$$\sum_{x_1 \leqq v_i \leqq x_2} p_i = x_2 - x_1 .$$

This is equivalent to

$$\sum_{0 \leqq v_i \leqq v} p_i = v, \qquad 0 \leqq v \leqq 1.$$

The deviation from uniformity can be measured by integrating the square of the difference between the left and right sides,

$$\int_0^1 \left(\sum_{0 \leqq v_i \leqq v} p_i - v \right)^2 dv.$$

If we define $F(v)$ by

$$F(v) = \sum_{0 \leqq v_i \leqq v} p_i ,$$

then the last formula becomes

$$\int_0^1 (F(v) - v)^2 \, dv.$$

Figure 11 gives a typical plot of $F(v)$ where $F(v) = 0$ for $v \leqq a$ and $F(v) = 1$ for $v \geqq b$. Now shifting $F(v)$ to the left or right [as long as neither a nor b is shifted out of the interval (0, 1)] does not change the essential shape of $F(v)$. It is reasonable to shift $F(v)$ to give the best approximation to v. We therefore define $E(s)$ to be the mean squared error when $F(v)$ is shifted by s; i.e.,

$$E(s) = \int_0^1 (F(v - s) - v)^2 \, dv \quad \text{for} \quad -a \leqq s \leqq 1 - b.$$

Then we redefine the deviation from uniformity by

$$E = \min_s E(s).$$

We derive the minimizing shift in the following lemma.

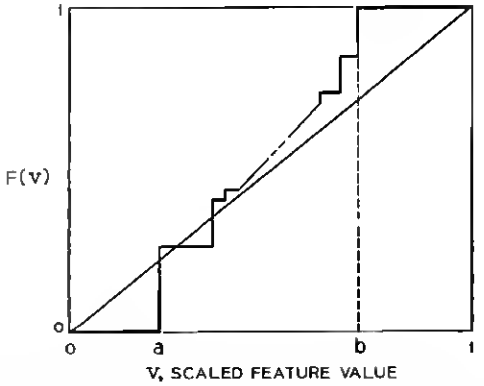

Fig. 11—A graph of a hypothetical $F(v)$, the cumulative distribution of P_i (the normalized weights) versus v (the scaled feature value). No individual has a scaled feature value less than a or greater than b.

Lemma: Let

$$E(s) = \int_0^1 (F(v - s) - v)^2 \, dv,$$

and let

$$e = \int_0^1 v \, dF(v).$$

Then for $-a < s < 1 - b$, $E(s)$ *is minimized for*

$$s = \begin{cases} -a & \text{if} \quad \tfrac{1}{2} - e \leqq -a \\ \tfrac{1}{2} - e & \text{if} \quad -a < \tfrac{1}{2} - e \leqq 1 - b \\ 1 - b & \text{if} \quad 1 - b < \tfrac{1}{2} - e. \end{cases}$$

Proof: To avoid needless mathematical complexity, let us suppose that $F(v)$ is a differentiable function. Then

$$E'(s) = -2 \int_0^1 (F(v - s) - v) F'(v - s) \, dv$$

$$= -F^2(1 - s) + F^2(-s) + 2 \int_0^1 v \, dF(v - s).$$

The first term is -1 since $b \leqq 1 - s$. The second term is 0 since $-s \leqq a$. The third term is easily shown to be $2(e + s)$ by using the substitution $u = v - s$ and the facts that $-s \leqq a$ and $b \leqq 1 - s$. Thus

$$E'(s) = 2(e + s - \tfrac{1}{2}).$$

$E(s)$ is minimized by having $e + s$ as close to $\frac{1}{2}$ as possible since $E'(s)$ is negative (positive) if $e + s$ lies to the left (right) of $\frac{1}{2}$.

APPENDIX B

Population Size and Identification Speed

We wish to show that the number of steps in the identification process grows at most logarithmically with P, the population size. More precisely, let r_k denote the rank of the target after the subject has given the kth feature value. It will be shown that under reasonable assumptions, given below, the expected value of r_k, for large k, satisfies

$$E(r_k) < P \cdot c^k = \exp\left(\ln P - k \ln \frac{1}{c}\right),$$

where $0 \leqq c < 1$, and c is a function of the distributions of the official values and the subject's errors in judgment. Thus, to achieve a given expected rank, the number of steps, k, need grow no faster than $\ln P$.

While we believe that these several assumptions lead to a reasonable model of our experiment, we expect them to provide only a qualitative indication of the growth of rank with population size. A quantitative analytical model is unobtainable at this point since the data we have are insufficient to extract the necessary statistical parameters. The assumptions are as follows:

Each of the P individuals in the population can be considered to be a vector $i = (i_1, i_2, \cdots)$ whose components are the official feature-values. We assume that these feature values are independent, identically distributed random variables and that the individuals are independent vectors. The subject describes the features of the target $t = (t_1, t_2, \cdots)$, and his judgments of the features are in error by $e_1, e_2, \cdots$. We assume that the errors are independent, identically distributed random variables.

By convention, the components of each vector are ordered in the sequence in which they are described by the subject.

Using the above notation and our definition of weight, the target has weight

$$\exp\left(-\sum_{j=1}^{k} | e_j |\right)$$

while an individual, i, has weight

$$\exp\left(-\sum_{j=1}^{k} | t_j + e_j - i_j |\right).$$

If we define

$$x_j(t, i) = | t_i + e_j - i_j | - | e_i | ,$$

then i's weight is larger than t's weight if

$$\sum_{j=1}^{k} x_j(t, i) < 0.$$

Let

$$s_k(t, i) = \begin{cases} 1 & \text{if } \sum_{j=1}^{k} x_j(t, i) < 0 \\ 0 & \text{otherwise.} \end{cases}$$

Define r_k, the rank of t, as the number of individuals with weight larger than t's weight. We then have

$$r_k = \sum_{i, i \neq t} s_k(t, i).$$

The expected value of r_k is

$$E(r_k) = E \sum_{i, i \neq t} s_k(t, i).$$

In $s_k(t, i)$, each summand $x_k(t, i)$ has, for fixed t, a distribution which clearly is a function of t. However, we are taking an expectation over all targets and populations. Thus, in this context, the $x_k(t, i)$ (for $t \neq i$) are independent, identically distributed variables since the t's, as well as the i's and e's, are independent, identically distributed variables. Hence

$$\begin{aligned} E(r_k) &= (P - 1)E(s_k(t, i)) = (P - 1) \Pr \{s_k(t, i) = 1\} \\ &= (P - 1) \Pr \left\{ \sum_{i=1}^{k} x_j(t, i) < 0 \right\}. \end{aligned}$$

Let the x_k's have common mean m and standard deviation σ. We apply the Central Limit Theorem to the last probability to obtain

$$\begin{aligned} \Pr \left\{ \sum_{j=1}^{k} x_j(t, i) < 0 \right\} &= \Pr \left\{ \frac{\sum_{j=1}^{k} x_j(t, i) - km}{\sqrt{k}\,\sigma} < -\sqrt{k}\,\frac{m}{\sigma} \right\} \\ &\sim \Phi(-\sqrt{k}\, m/\sigma) \end{aligned}$$

where Φ is the cumulative normal distribution. For large values of $\sqrt{k}\, m/\sigma$, the asymptotic formula* for Φ gives

* As $x \to \infty$, $\Phi(-x) \sim \frac{1}{\sqrt{2\pi}} \frac{e^{-x^2/2}}{x}$.

$$E(r_k) \sim (P - 1) \frac{1}{\sqrt{2\pi}} \frac{e^{-km^2/2\sigma^2}}{\sqrt{k}\, m/\sigma}$$

$$< P(e^{-m^2/2\sigma^2})^k = Pc^k$$

for k sufficiently large.

REFERENCES

1. Goldstein, A. J., Harmon, L. D., and Lesk, A. B., "Identification of Human Faces," Proc. IEEE, *59*, No. 5(1971), pp. 748–760.
2. Kingston, C. R., "Problems in Semi-Automated Fingerprint Classification," in *Proceedings of the First National Symposium on Law Enforcement Science and Technology*, Washington, D. C.: Thompson Book Co., 1967, pp. 449–457.
   Kingston, C. R., and Rudie, D. D., "Fingerprint Classification Methods Study—Second Status Report," N. Y. State Identification and Intelligence System/System Development Corp., October 1966.
3. Sitar, E. J., "A Handwriter Identification System," unpublished work.
4. Salton, G., and Lesk, M. E., "The SMART Automatic Document Retrieval System—an Illustration," Comm. ACM, *8*, No. 6 (1965), pp. 391–398.
5. Lesk, A. B., "An Interactive Program for Identification of Multi-Dimensional Objects," unpublished work.